SINAI at CLEF 2006 Ad Hoc Robust Multilingual Track: Query Expansion Using the Google Search Engine

Fernando Martínez-Santiago, Arturo Montejo-Ráez,
Miguel Á. García-Cumbreras, and L. Alfonso Ureña-López

SINAI Research Group. Computer Science Department. University of Jaén. Spain
{dofer,amontejo,magc,laurena}@ujaen.es

Abstract. This year, we have participated in the Ad-Hoc Robust Multilingual track with the aim of evaluating two important issues in Cross-Lingual Information Retrieval (CLIR) systems. This paper first describes the method applied for query expansion in a multilingual environment by using web search results provided by the Google engine in order to increase retrieval robustness. Unfortunately, the results obtained are disappointing. The second issue reported alludes to the robustness of several common merging algorithms. We have found that 2-step RSV merging algorithms perform better than others algorithms when evaluating using geometric average[1].

1 Introduction

Robust retrieval has been a task in the TREC evaluation forum [1]. One of the best systems proposed involves query expansion through web assistance [4,3,2]. We have followed the approach of Kwok and his colleagues and applied it to robust multilingual retrieval.

Pseudo-relevance feedback has been used traditionally to generate new queries from the results obtained from a given source query. Thus, the search is launched twice: one for obtaining first relevant documents from which new query terms are extracted, and a second turn to obtain final retrieval results. This method has been found useful to solve queries producing small result sets, and is a way to expand queries with new terms that can widen the scope of the search. But pseudo-relevance feedback is not that useful when queries are so difficult that very few or no documents are obtained at a first stage (the so-called weak queries). In that case, there is a straighforward solution: use a different and richer collection to expand the query. Here, the Internet plays a central role: it is a huge amount of web pages where virtually any query, no matter how difficult it is, may be related to some subset of those pages. This approach has obtained

[1] This work has been supported by the Spanish Government (MCYT) with grant TIC2003-07158-C04-04 and the RFC/PP2006/Id_514 granted by the University of Jaén.

C. Peters et al. (Eds.): CLEF 2006, LNCS 4730, pp. 119–126, 2007.